

---

# Introduction to Grid Computing

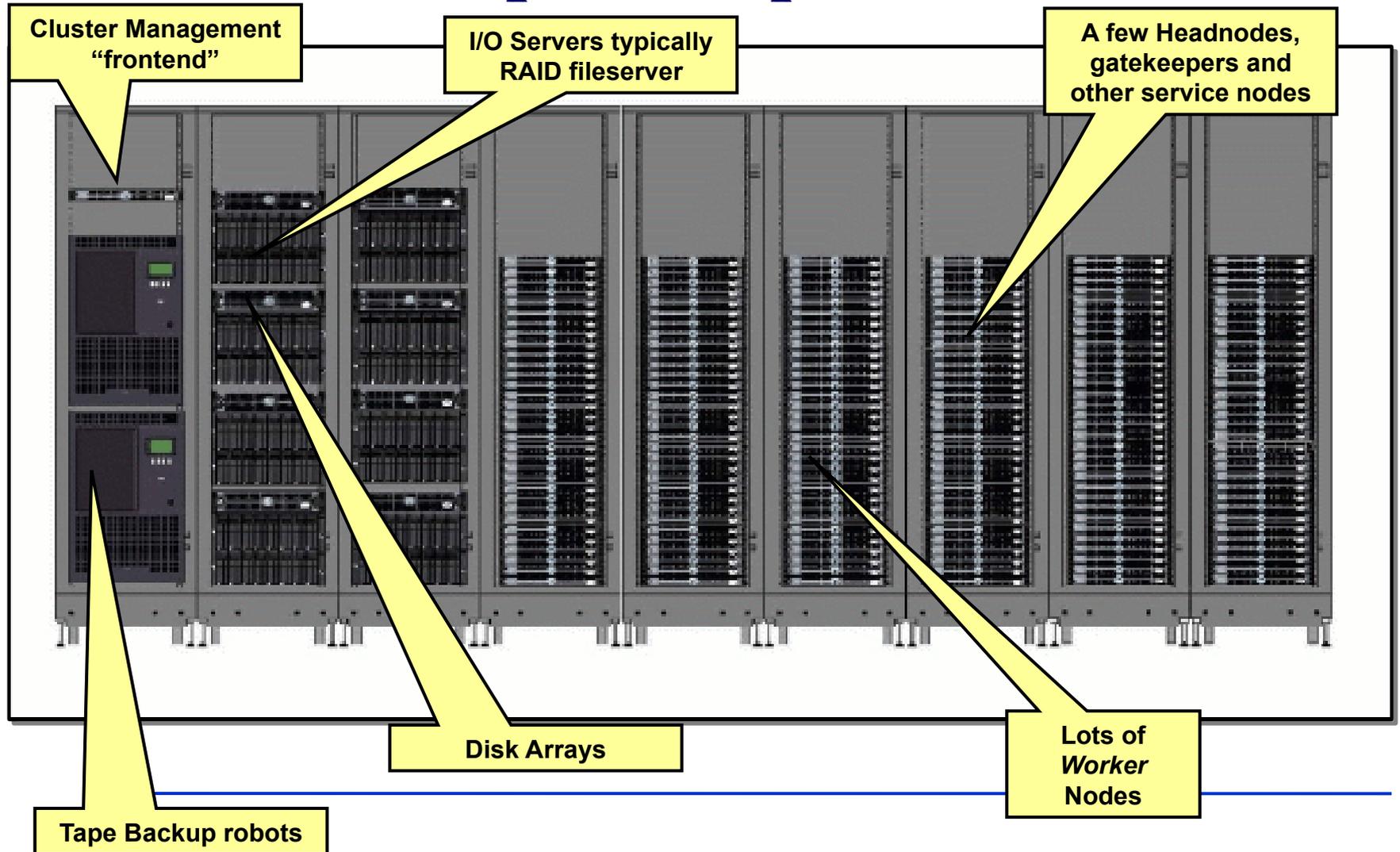
---

Grid School Workshop – Module 1

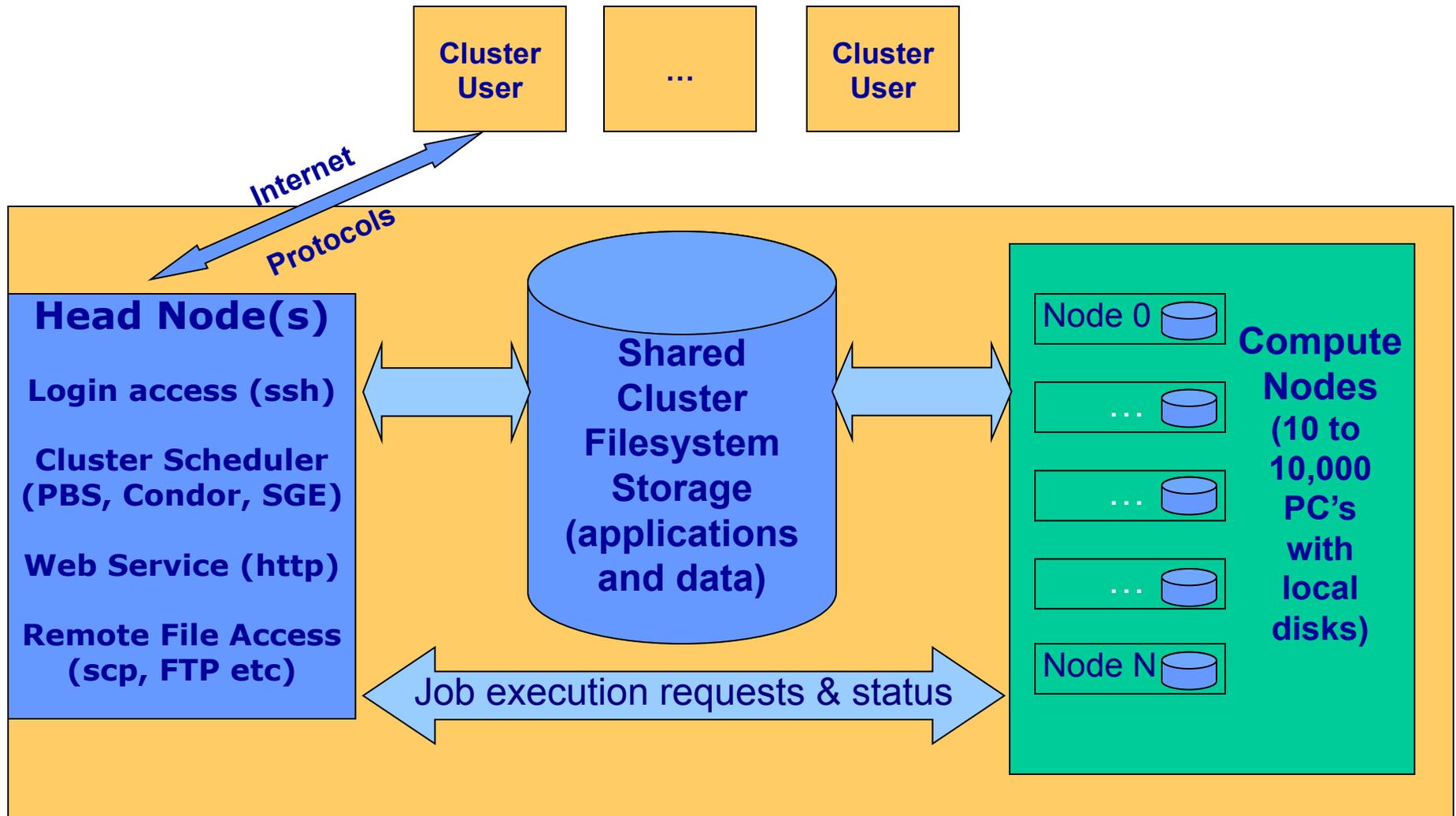


Open Science Grid

# Computing “Clusters” are today’s Supercomputers



# Cluster Architecture



# Scaling up Science: Citation Network Analysis in Sociology

## SHATTERPROOF MADS-box genes control seed dispersal in *Arabidopsis*

Sarah J. Lijlegren<sup>1</sup>, Gary S. Ditta<sup>1</sup>, Yuval Eshed<sup>1</sup>, Beth Savidge<sup>1</sup>, John L. Bowman<sup>1</sup> & Martin F. Yanofsky<sup>1</sup>

<sup>1</sup>Section of Cell and Developmental Biology, University of California at San Diego, La Jolla, California 92093-0116, USA

<sup>2</sup>Section of Plant Biology, University of California at Davis, Davis, California 95616, USA

Evolution of seed dispersal and dispersal of seeds, is a key trait in many flowering plants. Seed dispersal in *Arabidopsis* occurs primarily by a process called fruit dehiscence, or pod shatter. Few studies have focused on identifying genes that regulate this process, in spite of the agronomic value of controlling seed dispersal in crop plants such as canola<sup>3,4</sup>. Here we show that the closely related SHATTERPROOF (SHP) and SHATTERPROOF2 (SHP2) MADS-box genes are required for fruit dehiscence in *Arabidopsis*.

Keywords: fruit dehiscence, pod shatter, MADS-box genes, seed dispersal, *Arabidopsis*

Introduction: Seed dispersal is a key trait in many flowering plants. Seed dispersal in *Arabidopsis* occurs primarily by a process called fruit dehiscence, or pod shatter. Few studies have focused on identifying genes that regulate this process, in spite of the agronomic value of controlling seed dispersal in crop plants such as canola<sup>3,4</sup>. Here we show that the closely related SHATTERPROOF (SHP) and SHATTERPROOF2 (SHP2) MADS-box genes are required for fruit dehiscence in *Arabidopsis*.

Results: We show that SHP and SHP2 are required for fruit dehiscence in *Arabidopsis*. SHP and SHP2 are expressed in the fruit wall and are required for the development of the fruit wall. SHP and SHP2 are required for the development of the fruit wall. SHP and SHP2 are required for the development of the fruit wall.

Discussion: SHP and SHP2 are required for fruit dehiscence in *Arabidopsis*. SHP and SHP2 are expressed in the fruit wall and are required for the development of the fruit wall. SHP and SHP2 are required for the development of the fruit wall.

Conclusion: SHP and SHP2 are required for fruit dehiscence in *Arabidopsis*. SHP and SHP2 are expressed in the fruit wall and are required for the development of the fruit wall. SHP and SHP2 are required for the development of the fruit wall.

References: 1. ... 2. ... 3. ... 4. ...

Supplementary Information: ...

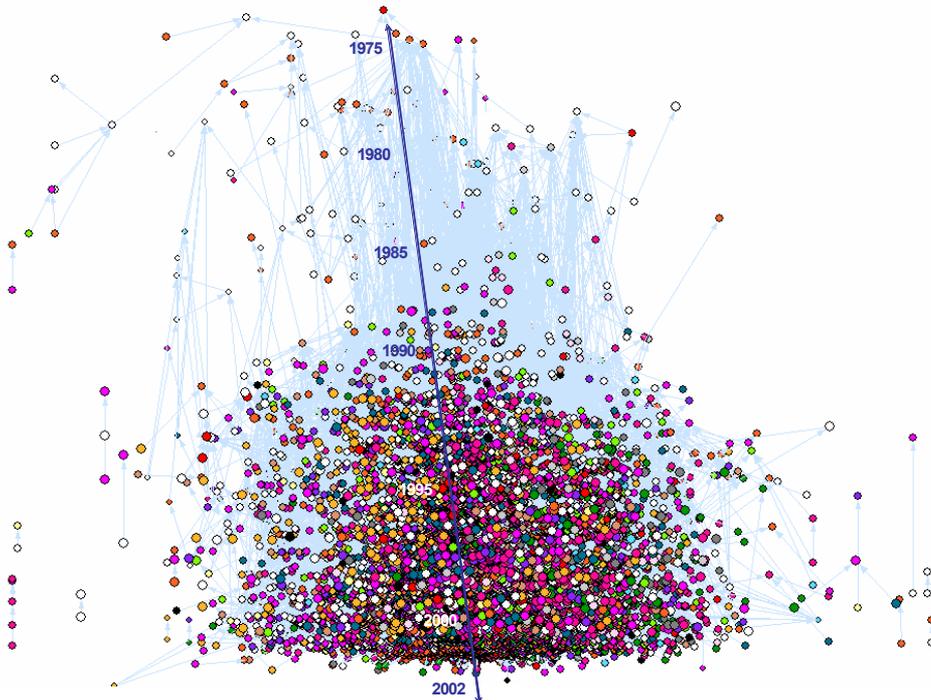
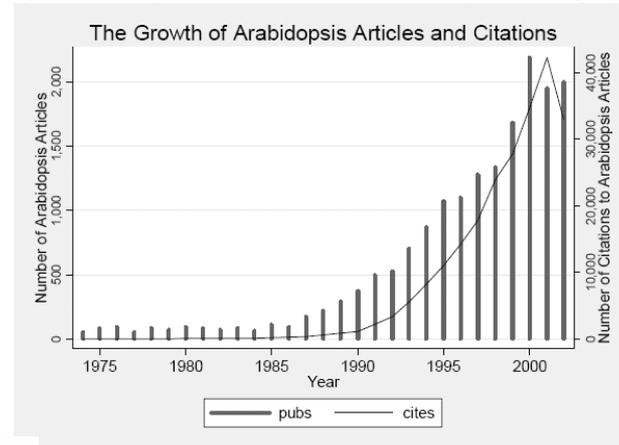
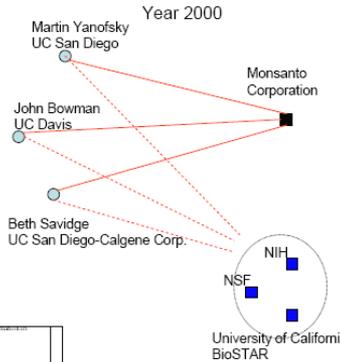
Correspondence: ...

Additional Information: ...

Supplementary Information: ...

Correspondence: ...

Additional Information: ...



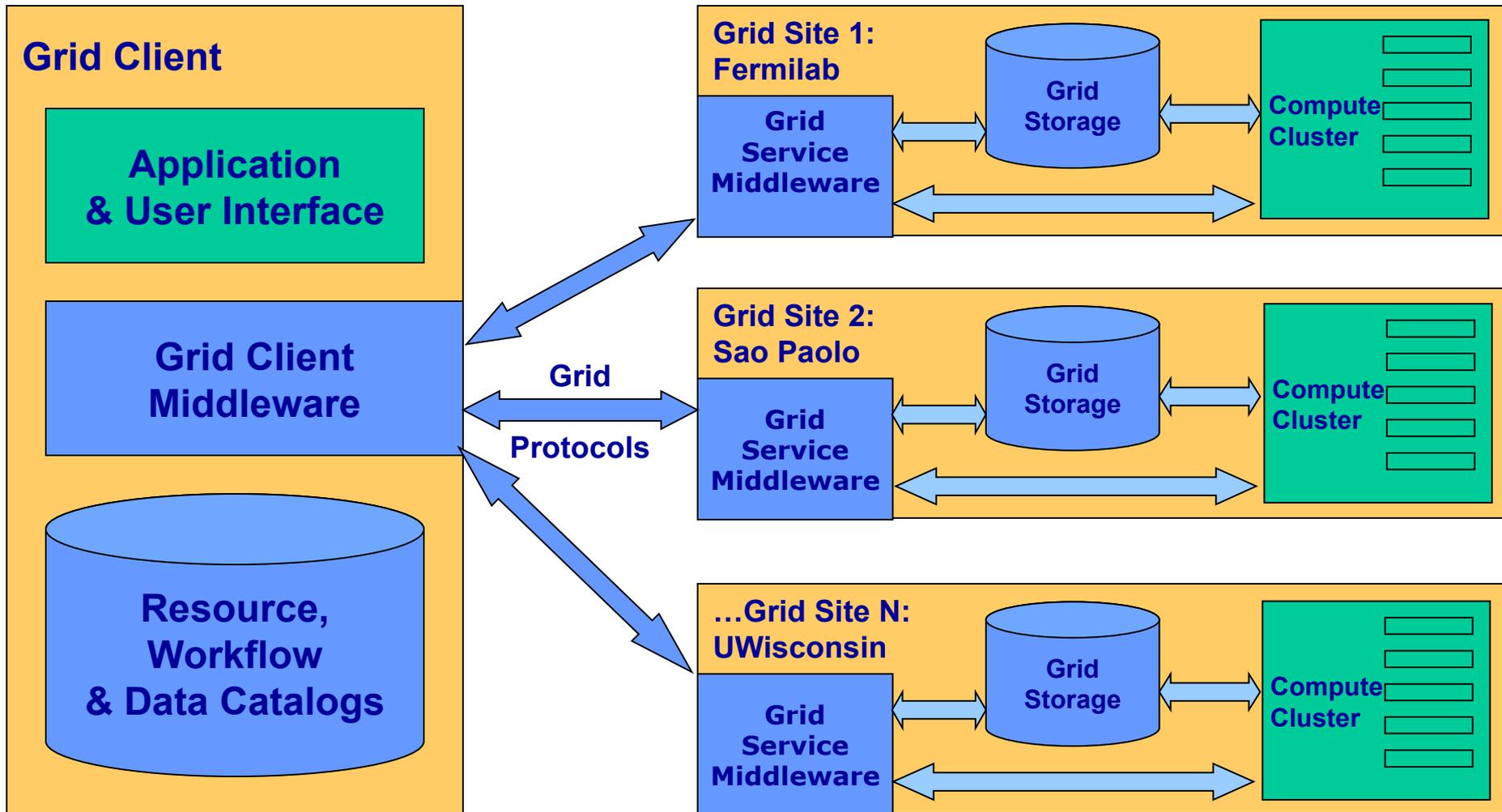
- miscellaneous
- transcription factors / morphogens
- receptors
- phosphosphorylation cascades
- organogenesis
- terpenes, synthesis
- physical defense
- commercial disease resistance
- innate immunology
- nutrient metabolism and movement
- nutrient uptake
- genomics
- photosynthesis
- functional enzymatics
- protein isolation & characterization
- targeting / splicing
- tropisms

Work of James Evans,  
University of Chicago,  
Department of  
Sociology

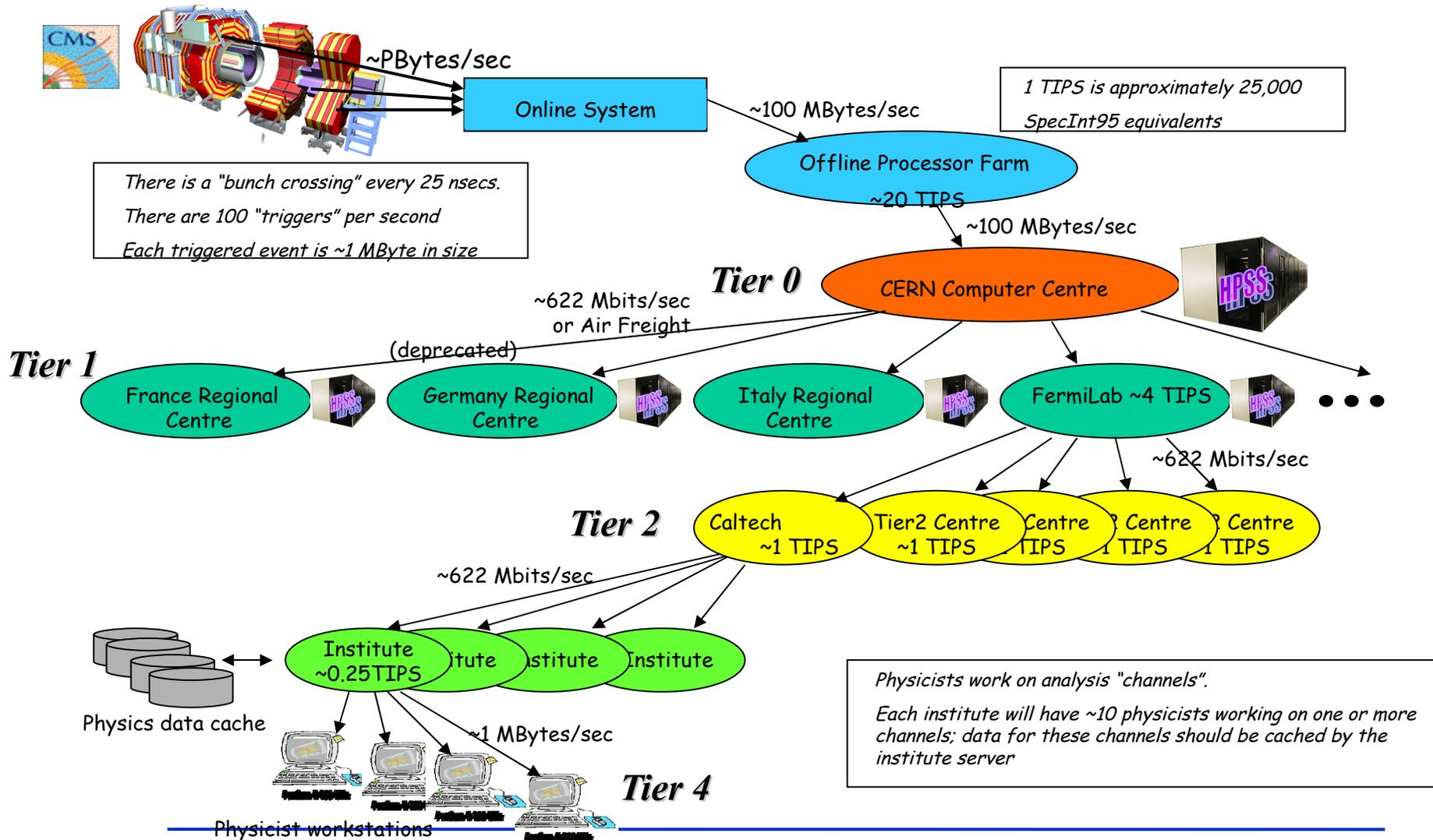
# Scaling up the analysis

- Query and analysis of 25+ million citations
- Work started on desktop workstations
- Queries grew to month-long duration
- With data distributed across  
U of Chicago TeraPort **cluster**:
  - 50 (faster) CPUs gave 100 X speedup
  - Many more methods and hypotheses can be tested!
- Higher *throughput* and *capacity* enables *deeper analysis* and *broader community access*.

# Grids consist of distributed clusters

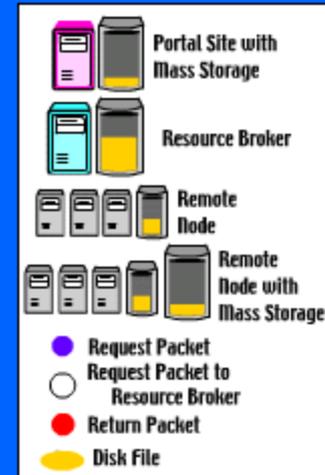


# Initial Grid driver: High Energy Physics



# Grids Provide Global Resources To Enable e-Science

## World GRID Computing

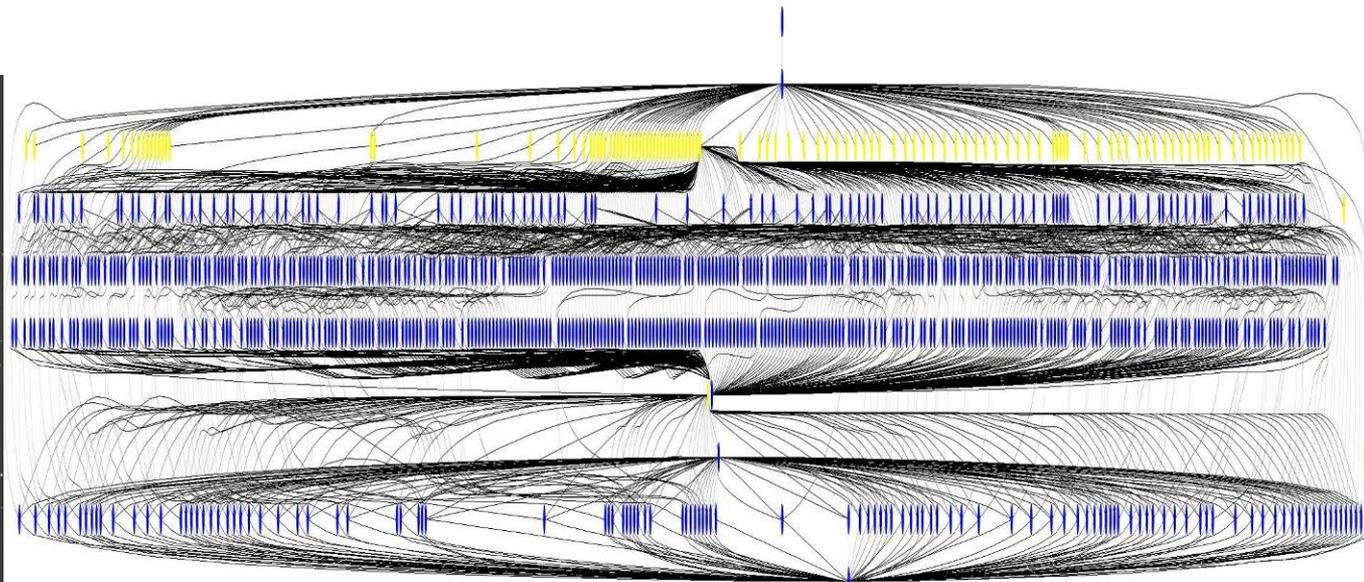


# Grids can process vast datasets.

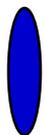
- Many HEP and Astronomy experiments consist of:
  - Large datasets as inputs (find datasets)
  - “Transformations” which work on the input datasets (process)
  - The output datasets (store and publish)
- The emphasis is on the sharing of these large datasets
- *Workflows of independent program can be parallelized.*



Mosaic of M42 created on TeraGrid



= Data  
Transfer



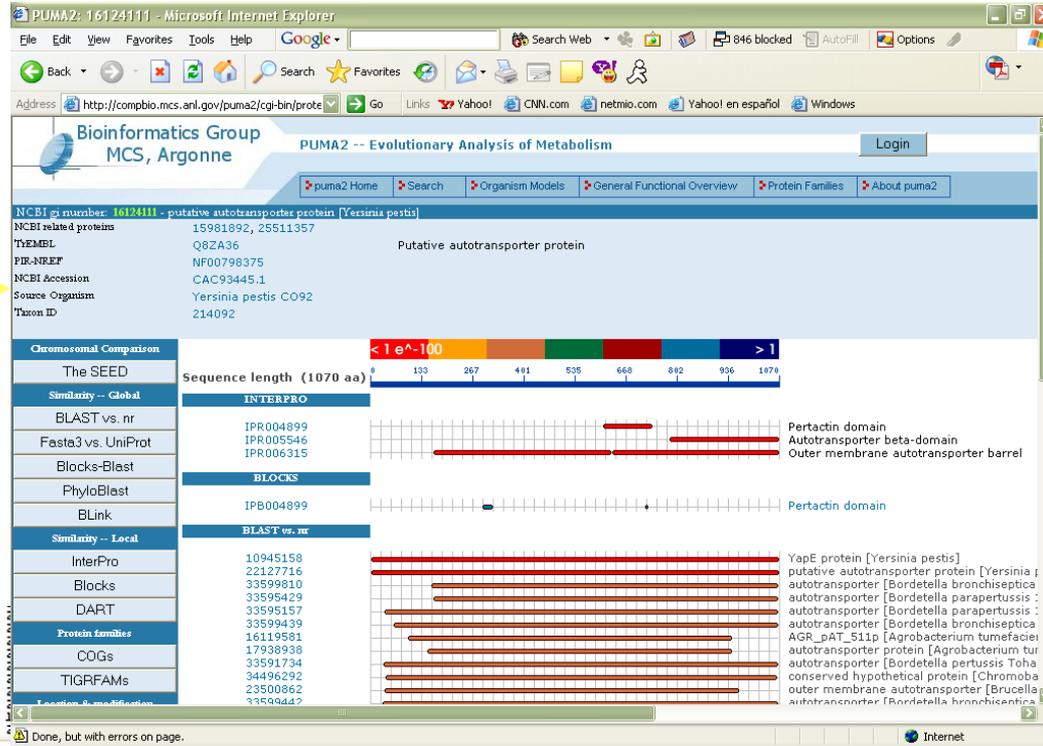
= Compute  
Job

Montage Workflow: ~1200 jobs, 7 levels  
NVO, NASA, ISI/Pegasus - Deelman et al.

# PUMA: Analysis of Metabolism

## PUMA Knowledge Base

Information about proteins analyzed against ~2 million gene sequences



gi 23499780 gn REF_tigr BRA0013	gi 16080253 ref NP_391080.1	44.27	253	131	1	15	257	8
gi 23499780 gn REF_tigr BRA0013	gi 23098409 ref NP_691875.1	43.48	253	133	2	16	258	5
gi 23499780 gn REF_tigr BRA0013	gi 48837187 ref ZP_00294182.1	44.92	256	125	2	14	256	7
gi 23499780 gn REF_tigr BRA0013	gi 52005400 gb AA025942.1	44.75	257	125	2	15	258	3
gi 23499780 gn REF_tigr BRA0013	gi 48864015 ref ZP_00317908.1	44.49	245	134	1	13	257	5
gi 23499780 gn REF_tigr BRA0013	gi 30348891 gb AA028934.1	39.53	253	138	3	18	257	5
gi 23499780 gn REF_tigr BRA0013	gi 19655222 gb AA093939.1	40.64	251	138	1	17	256	10
gi 23499780 gn REF_tigr BRA0013	gi 2735808 gb AA007757.1	43.03	251	130	4	18	256	11
gi 23499780 gn REF_tigr BRA0013	gi 12597924 gb AA035899.2	46.70	182	96	1	62	243	5
gi 23499780 gn REF_tigr BRA0013	gi 46363318 ref ZP_00236079.1	39.58	240	135	2	14	253	6

REF_tigr BRA0013	gi 39933731 ref NP_946007.1	34.90	255	e-33 142.5
REF_tigr BRA0013	gi 48782600 ref ZP_00279106.1	35.92	245	e-32 141.4
REF_tigr BRA0013	gi 41407534 ref NP_960370.1	36.09	266	e-32 140.2
REF_tigr BRA0013	gi 48851585 ref ZP_00305793.1	32.39	247	e-32 139.4
REF_tigr BRA0013	gi 15966306 ref NP_386659.1	36.50	263	e-32 139.0
REF_tigr BRA0013	gi 17548526 ref NP_521866.1	36.36	264	e-31 137.9

gi 23499780 gn REF_tigr BRA0013	gi 51891730 ref VP_074421.1	38.87	247	136	7	18	256	1	2403.4	e-30 133.7
gi 23499780 gn REF_tigr BRA0013	gi 145881 gb AA23739.1	33.87	246	147	3	13	253	3	2404.4	e-30 133.3
gi 23499780 gn REF_tigr BRA0013	gi 25029334 ref NP_739388.1	35.20	250	147	4	15	256	6	2485.7	e-30 132.0
gi 23499780 gn REF_tigr BRA0013	gi 21230953 ref NP_636732.1	36.52	257	138	6	12	255	5	2545.7	e-30 132.0
gi 23499780 gn REF_tigr BRA0013	gi 46314029 ref ZP_00246316.1	33.86	254	153	2	12	259	3	2485.7	e-30 132.0
gi 23499780 gn REF_tigr BRA0013	gi 41406852 ref NP_959688.1	35.61	238	149	2	16	253	2	2309.8	e-30 132.1
gi 23499780 gn REF_tigr BRA0013	gi 15644471 ref NP_229523.1	35.69	255	144	5	12	256	2	2469.8	e-30 132.1
gi 23499780 gn REF_tigr BRA0013	gi 23470090 ref ZP_00125423.1	35.20	250	145	4	12	257	4	2439.8	e-30 132.1
gi 23499780 gn REF_tigr BRA0013	gi 24935279 gb AA064837.1	34.63	257	146	4	12	257	4	2499.9	e-30 132.1
gi 23499780 gn REF_tigr BRA0013	gi 48647651 ref ZP_0030215.1	36.05	258	145	9	12	257	4	2531.2	e-29 131.7

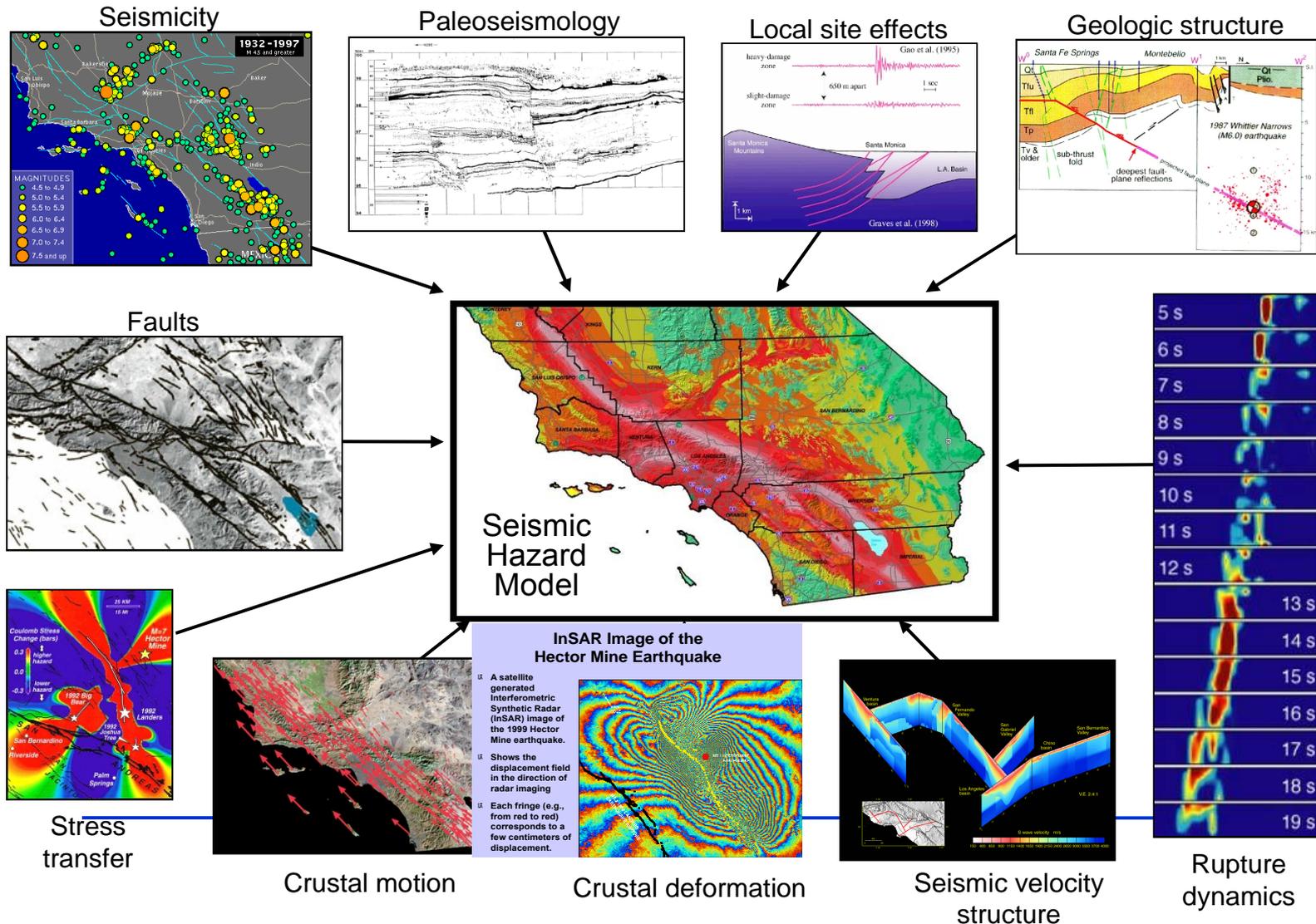
Natalia Maltsev et al.

<http://compbio.mcs.anl.gov/puma2>

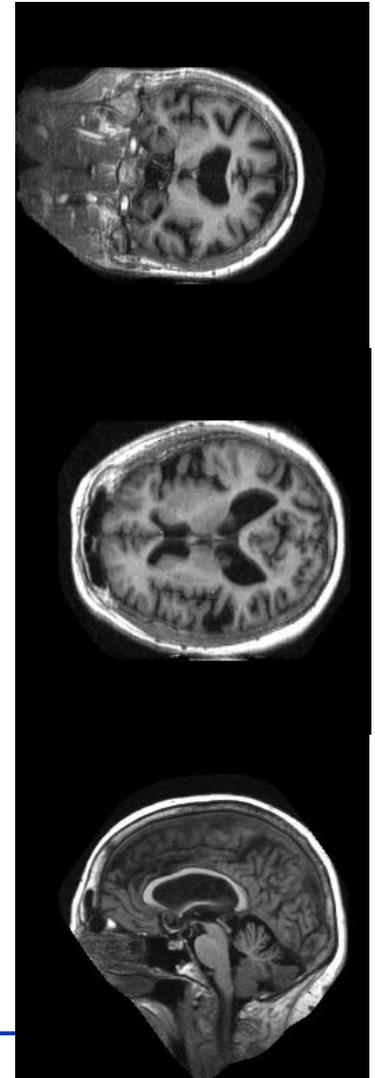
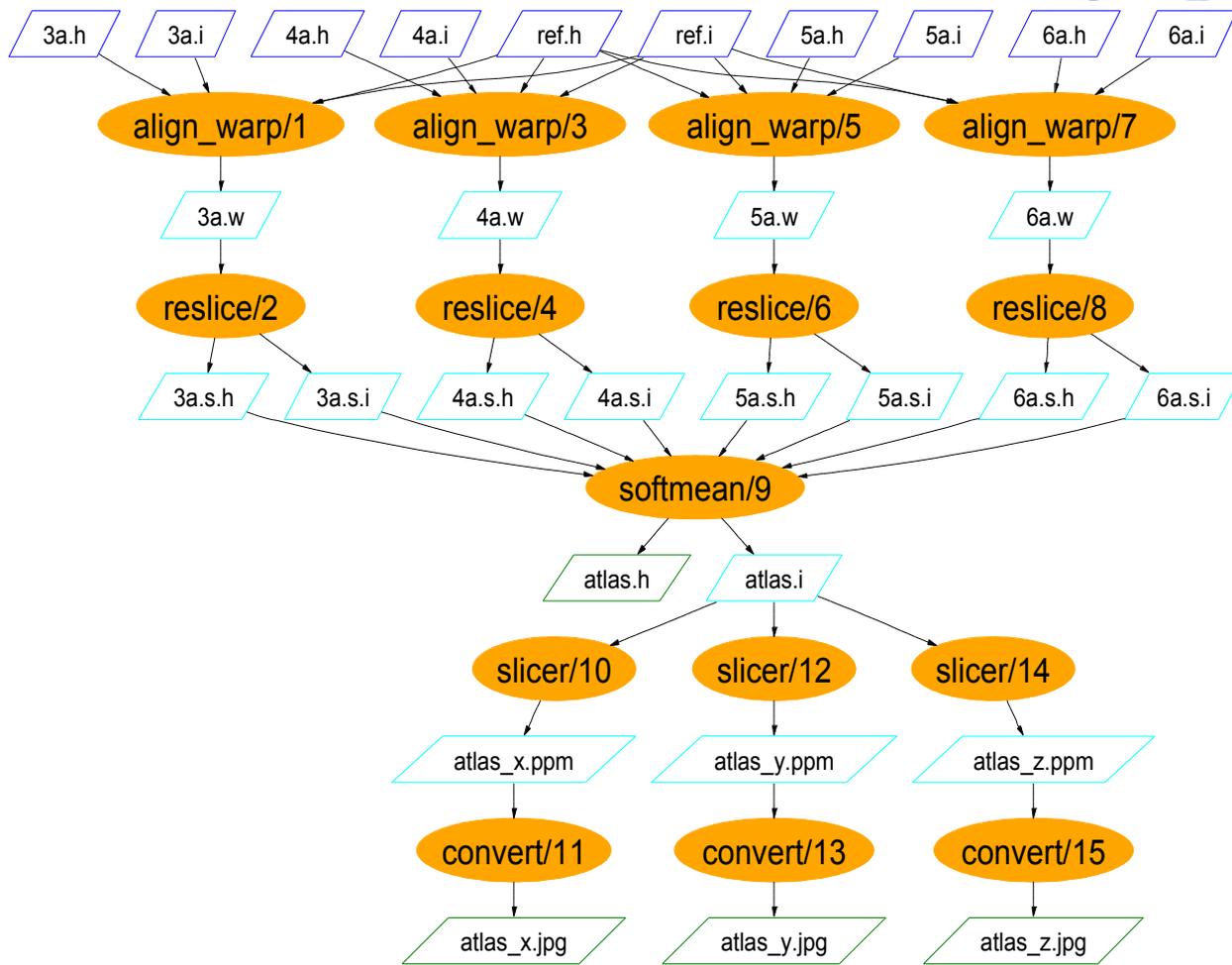
## Analysis on Grid

Involves millions of BLAST, BLOCKS, and other processes

# Mining Seismic data for hazard analysis (Southern Calif. Earthquake Center).



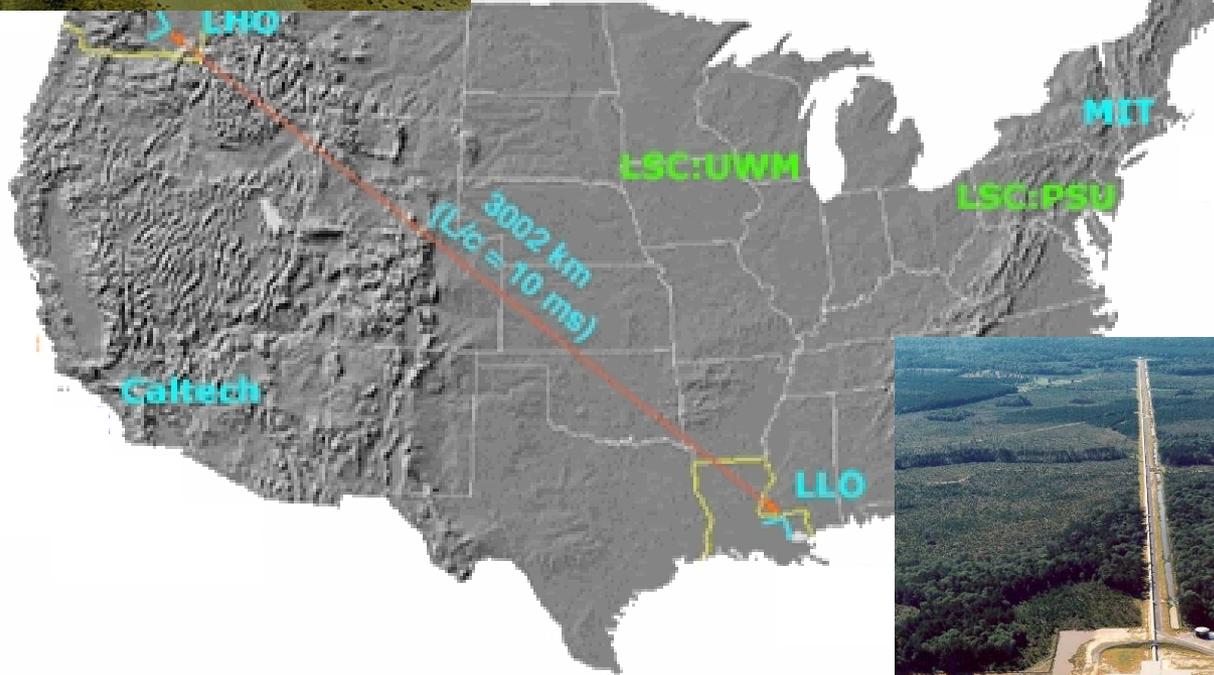
# A typical workflow pattern in image analysis runs many filtering apps.





# The Globus-Based LIGO Data Grid

LIGO Gravitational Wave Observatory



Replicating  $>1$  Terabyte/day to 8 sites  
>40 million replicas so far  
MTBF = 1 month

# Virtual Organizations

- Groups of organizations that use the Grid to share resources for specific purposes
- Support a single community
- Deploy compatible technology and agree on working policies
  - Security policies - difficult
- Deploy different network accessible services:
  - Grid Information
  - Grid Resource Brokering
  - Grid Monitoring
  - Grid Accounting



---

# Ian Foster's Grid Checklist

- A Grid is a system that:
  - Coordinates resources that are not subject to centralized control
  - Uses standard, open, general-purpose protocols and interfaces
  - Delivers non-trivial qualities of service

# The Grid Middleware Stack *(and course modules)*

**Grid Application (M5)**  
(often includes a *Portal*)

**Workflow system (explicit or *ad-hoc*) (M6)**

**Job  
Management (M2)**

**Data  
Management (M3)**

**Grid Information  
Services (M5)**

**Grid Security Infrastructure (M4)**

**Core Globus Services (M1)**

**Standard Network Protocols and *Web Services* (M1)**

# Globus and Condor play key roles

- Globus Toolkit provides the base middleware
  - Client tools which you can use from a command line
  - APIs (scripting languages, C, C++, Java, ...) to build your own tools, or use direct from applications
  - Web service interfaces
  - Higher level tools built from these basic components, e.g. Reliable File Transfer (RFT)
- Condor provides both client & server scheduling
  - In grids, Condor provides an agent to queue, schedule and manage work submission

# Grid architecture is evolving to a Service-Oriented approach.

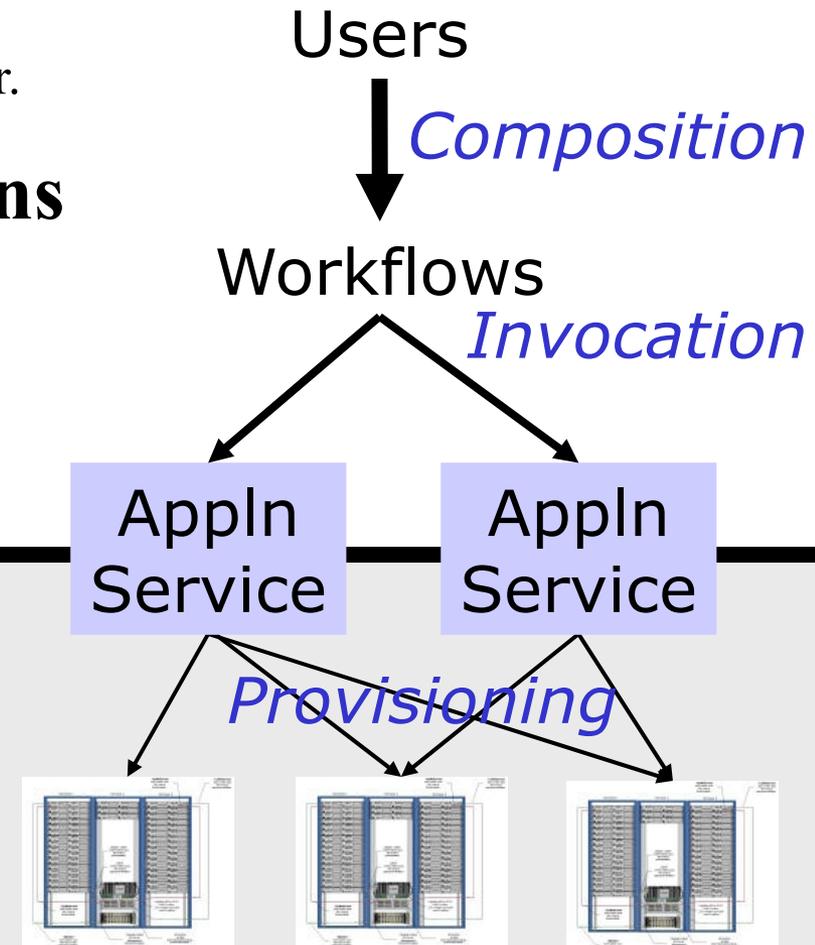
*...but this is beyond our workshop's scope.*  
See "Service-Oriented Science" by Ian Foster.

## ■ Service-oriented **applications**

- Wrap applications as services
- Compose applications into workflows

## ■ Service-oriented **Grid infrastructure**

- Provision physical resources to support application workloads



# Local Resource Manager: a batch scheduler for running jobs on a computing cluster

- Popular LRMs include:
  - PBS – Portable Batch System
  - LSF – Load Sharing Facility
  - SGE – Sun Grid Engine
  - Condor – Originally for cycle scavenging, Condor has evolved into a comprehensive system for managing computing
- LRMs execute on the cluster's *head node*
- Simplest LRM allows you to “fork” jobs quickly
  - Runs on the head node (*gatekeeper*) for fast utility functions
  - No queuing (but this is emerging to “throttle” heavy loads)
- In GRAM, each LRM is handled with a “job manager”

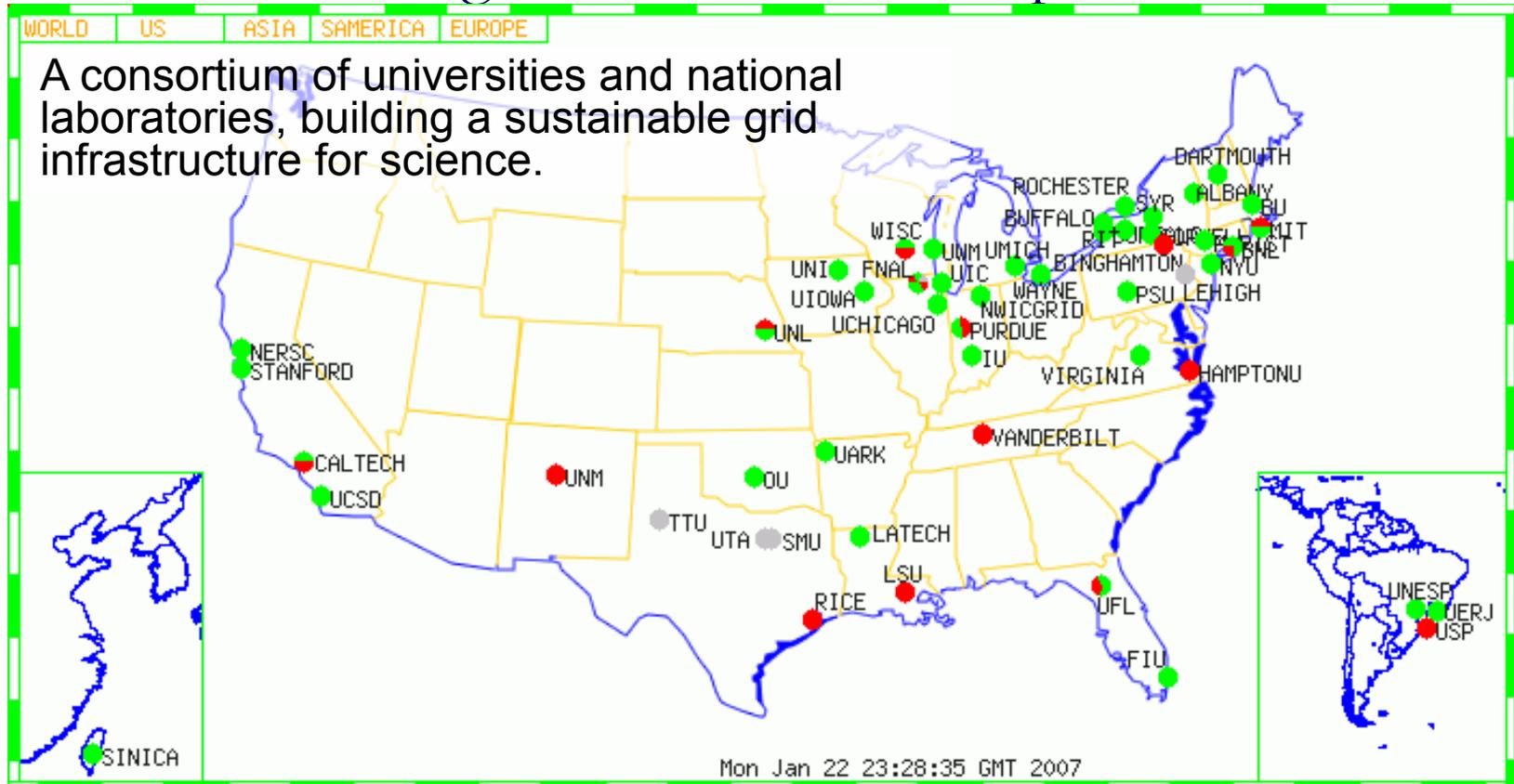
# Grid security is a crucial component

- Problems being solved might be sensitive
- Resources are typically valuable
- Resources are located in distinct administrative domains
  - Each resource has own policies, procedures, security mechanisms, etc.
- Implementation must be broadly available & applicable
  - Standard, well-tested, well-understood protocols; integrated with wide variety of tools

# Grid Security Infrastructure - GSI

- Provides secure communications for all the higher-level grid services
- Secure *Authentication* and *Authorization*
  - Authentication ensures you *are* whom you claim to be
    - *ID card, fingerprint, passport, username/password*
  - Authorization controls what you are permitted to *do*
    - *Run a job, read or write a file*
- GSI provides Uniform Credentials
- Single Sign-on
  - User authenticates once – then can perform many tasks

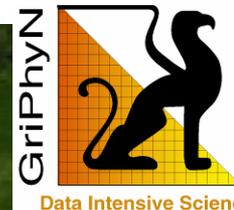
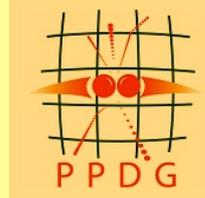
# Open Science Grid (OSG) provides shared computing resources, benefiting a broad set of disciplines



- OSG incorporates advanced networking and focuses on general services, operations, end-to-end performance
- Composed of a large number (>50 and growing) of shared computing facilities, or “sites”

# Open Science Grid

- 50 sites (15,000 CPUs) & growing
- 400 to >1000 concurrent jobs
- Many applications + CS experiments; includes long-running production operations
- Up since October 2003; few FTEs central ops



[www.opensciencegrid.org](http://www.opensciencegrid.org)

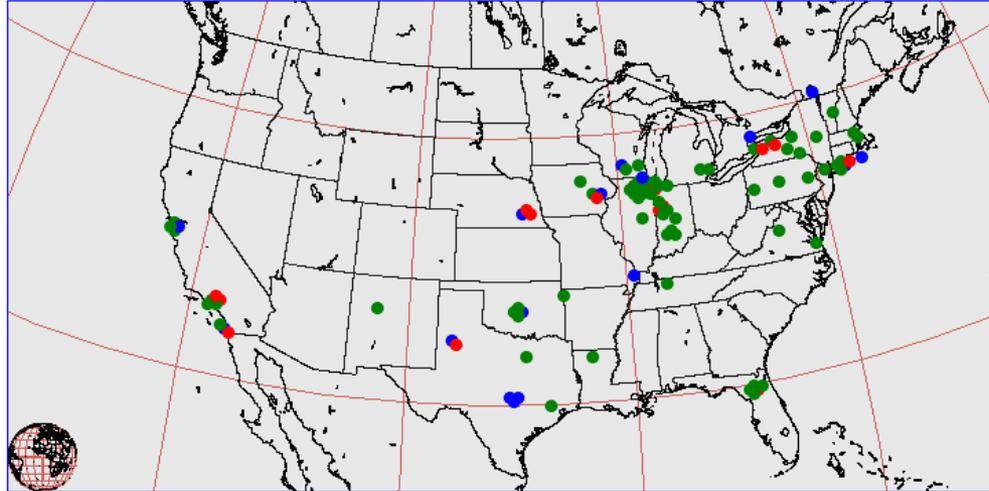
TeraGrid provides vast resources via a number of huge computing facilities.



---

To efficiently use a Grid, you must locate and monitor its resources.

- Check the availability of different grid sites
- Discover different grid services
- Check the status of “jobs”
- Make better scheduling decisions with information maintained on the “health” of sites



### Virtual Organization Selection

<input checked="" type="button" value="All"/>	<input type="button" value="CDF"/>	<input type="button" value="CMS"/>	<input type="button" value="CompBioGrid"/>	<input type="button" value="DES"/>	<input checked="" type="button" value="DOSAR"/>	<input type="button" value="DZero"/>	<input type="button" value="Engage"/>	<input type="button" value="Fermilab"/>	<input type="button" value="fMRI"/>	<input type="button" value="GADU"/>
	<input type="button" value="geant4"/>	<input type="button" value="GLOW"/>	<input type="button" value="GPN"/>	<input type="button" value="GRASE"/>	<input type="button" value="GridChem"/>	<input type="button" value="GridEx"/>	<input type="button" value="GROW"/>	<input type="button" value="i2u2"/>	<input type="button" value="iVDGL"/>	<input type="button" value="LIGO"/>
<input type="button" value="mariachi"/>	<input type="button" value="MIS"/>	<input type="button" value="nanoHUB"/>	<input type="button" value="NWICG"/>	<input type="button" value="Ops"/>	<input type="button" value="OSG"/>	<input type="button" value="OSGEDU"/>	<input type="button" value="SDSS"/>	<input type="button" value="STAR"/>	<input type="button" value="USATLAS"/>	

### Resources

Name	Gatekeeper	Type	Grid	Status	Last Test Date
<a href="#">BNL ATLAS 1</a>	gridgk01.racf.bnl.gov:2119	compute	OSG	PASS	2006-12-08 14:57:13
<a href="#">BNL ATLAS 2</a>	gridgk02.racf.bnl.gov:2119	compute	OSG	PASS	2006-12-08 14:58:43
<a href="#">BU ATLAS Tier2</a>	atlas.bu.edu:2119	compute	OSG	PASS	2006-12-08 15:00:44

# Conclusion: Why Grids?

- New approaches to inquiry based on
  - Deep analysis of huge quantities of data
  - Interdisciplinary collaboration
  - Large-scale simulation and analysis
  - Smart instrumentation
  - *Dynamically assemble the resources to tackle a new scale of problem*
- Enabled by access to resources & services without regard for location & other barriers

# Grids:

## Because Science needs community ...

- Teams organized around common goals
  - People, resource, software, data, instruments...
- With diverse membership & capabilities
  - Expertise in multiple areas required
- And geographic and political distribution
  - No location/organization possesses all required skills and resources
- Must adapt as a function of the situation
  - Adjust membership, reallocate responsibilities, renegotiate resources

---

Based on:

Grid Intro and Fundamentals Review

---

Dr. Gabrielle Allen

Center for Computation & Technology

Department of Computer Science

Louisiana State University

[gallen@cct.lsu.edu](mailto:gallen@cct.lsu.edu)